NIF – NLP Interchange Format

http://aksw.org/Projects/NIF

Sebastian Hellmann
AKSW, Universität Leipzig
Outline:

• NLP Interchange Format
• Use Cases
  – Integration of tools
  – Meaning Representation Language
  – Knowledge Extraction with SPARQL
  – Machine Learning
• Related Projects
Problem:
• Currently NLP software is organized in pipelines
• Integration is done „hard-wired“
  – For each tool and each framework an adapter has to be created (n * m)
• Difficult to aggregate output
• Difficult to exchange single components
Overview:

• NLP tools can be integrated via a common output format (Common pattern in Enterprise Application Integration)
• For each tool a wrapper needs to be created, that reads NIF and produces NIF
• The combination of tools can be adhoc, i.e. it is not a pipeline that needs to be configured
• Multi-layer and overlapping annotations are possible
• Ontologies provide interfaces for each layer and for applications
• First Challenge: Representing Strings in RDF
  • How to give a part of a document or text an identifier (URI)?
  • What properties can such URIs have?
Creating Knowledge out of Interlinked Data

Linked Data

The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data.

Like the web of hypertext, the web of data is constructed with documents on the web. However, unlike the web of hypertext, where links are relationships anchors in hypertext documents written in HTML, for data they links between arbitrary things described by RDF. The URIs identify any kind of object or concept. But for HTML or RDF, the same expectations apply to make the web grow:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs, so that they can discover more things.

Simple. In fact, though, a surprising amount of data isn't linked in 2006, because of problems with one or more of the steps. This article discusses solutions to these problems, details of implementation, and factors affecting choices about how you publish your data.

The four rules

I'll refer to the steps above as rules, but they are expectations of behavior. Breaking them does not destroy anything, but misses an opportunity to make data interconnected. This in turn limits the ways it can later be reused in unexpected ways. It is the unexpected re-use of information which is the value added by the web.

The first rule, to identify things with URIs, is pretty much understood by most people doing semantic web technology. If it doesn't use the universal URI set of symbols, we don't call it Semantic Web.

The second rule, to use HTTP URIs, is also widely understood. The only deviation has been, since the web started, a constant tendency for people to invent new URI schemes (and sub-schemes within the urn: scheme) such as LSIDs and handles and XRIs and DOIs and so on, for various reasons. Typically, these involve not wanting to commit to the established Domain Name System (DNS) for delegation of authority but to construct
## Example URIs for annotating „Semantic Web“

<table>
<thead>
<tr>
<th>Name</th>
<th>URI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offset (Start-End)</td>
<td>#14406-14418</td>
</tr>
<tr>
<td>Offset w/ recipe</td>
<td>#offset_Semantic+Web_14406-14418</td>
</tr>
<tr>
<td>Offset-Length w/ #</td>
<td>#14406-12</td>
</tr>
<tr>
<td>Offset-Length w/ /</td>
<td>/14406-12</td>
</tr>
<tr>
<td>Offset-Length w/ ?</td>
<td>?nif=14406-12</td>
</tr>
<tr>
<td>Context-Hash w/recipe</td>
<td>#hash_md5_4_12_Semantic+Web_abeb272fe2deadd2cd486c4cea6cddf1</td>
</tr>
<tr>
<td>Word w/o context</td>
<td>#word_Semantic+Web</td>
</tr>
<tr>
<td>Yee (Context)</td>
<td>#:words:we-dont-call-it-(Semantic Web).-The-second-rule,-to-use</td>
</tr>
<tr>
<td>Wilde and Dürst</td>
<td>#char=14406-12</td>
</tr>
<tr>
<td>LiveURL (Content-Based)</td>
<td>#8Semantic12+0x206A73ED</td>
</tr>
</tbody>
</table>
### NIF – NLP Interchange Format

<table>
<thead>
<tr>
<th></th>
<th>Uniq</th>
<th>Arb</th>
<th>XML</th>
<th>Trans</th>
<th>Addr</th>
<th>Self</th>
<th>Impl</th>
<th>Func</th>
</tr>
</thead>
<tbody>
<tr>
<td>Context-Hash (NIF)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Offset</td>
<td>++</td>
<td>++</td>
<td>-</td>
<td>- -</td>
<td>++</td>
<td>-</td>
<td>++</td>
<td>o</td>
</tr>
<tr>
<td>Offset w/ recipe (NIF)</td>
<td>++</td>
<td>++</td>
<td>+</td>
<td>- -</td>
<td>++</td>
<td>+</td>
<td>++</td>
<td>o</td>
</tr>
<tr>
<td>Yee (Context)</td>
<td>+</td>
<td>- -</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>o</td>
</tr>
<tr>
<td>Wilde and Dürst (Param)</td>
<td>++</td>
<td>++</td>
<td>+</td>
<td>- -</td>
<td>++</td>
<td>++</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>LiveURL (Content-Based)</td>
<td>- -</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>++</td>
<td>o</td>
</tr>
<tr>
<td>LiveURL (Position-Based)</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>- -</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>o</td>
</tr>
<tr>
<td>Wilde et.al. (Regex)</td>
<td>o</td>
<td>++</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>- -</td>
<td>++</td>
</tr>
</tbody>
</table>

Table 2.2: Comparison of URI recipes
• URIs are used to integrate output. RDF merges naturally, if the URIs are the same (or convertible using a certain recipe)
• Second challenge: Output of each layer is required to be stable.
  • Components and layers can be interchanged
  • Domain ontologies are needed to provide stable interfaces:
    – OLiA provides an ontological interface for morpho-syntax
      http://nachhalt.sfb632.uni-potsdam.de/owl/
    – DBpedia provides stable ids for Things
Creating Knowledge out of Interlinked Data

NIF – NLP Interchange Format
Creating Knowledge out of Interlinked Data

NIF – NLP Interchange Format

KAIST LOD2  17.8.2011

http://lod2.eu
Creating Knowledge out of Interlinked Data

NIF – NLP Interchange Format
Creating Knowledge out of Interlinked Data

Demo - Integration

• http://nlp2rdf.lod2.eu/annotator-stanford/NIFStemmer?input=My%20favorite%20actress%20is%20Natalie%20Portman!

• http://nlp2rdf.lod2.eu/annotator-stanford/NIFStanfordCore?input=My%20favorite%20actress%20is%20Natalie%20Portman!
Use Cases

• Use Cases
  – Integration of tools
  – Meaning Representation Language
  – Knowledge Extraction with SPARQL
  – Machine Learning
Use Case – Integration of tools
Use Case – Meaning Representation Language

- RDF makes data integration easy: URIref, LinkedData
- OWL is based on Description Logics (Guarded Fragment)
- Availability of open data sets (access and licence)
- Diverse serializations for annotations: XML, Turtle, RDFa+XHTML
- Scalable tool support (Databases, Reasoning)
Creating Knowledge out of Interlinked Data

Use Case - Meaning Representation Language

Explicit Meaning

Semantic Gap

Existing structured knowledge is selected, disambiguated and integrated

WSD connects top and bottom

Each NLP layer is augmented with linguistic background knowledge

Backbone ontology

Plain Text

Implicit Meaning

http://lod2.eu
• Classical approach:
  • POS tag / Dependency parser (e.g. Stanford)
  • create a rule/pattern language to extract knowledge
Use Case – Knowledge Extraction with SPARQL

# Example:
# A fish is any aquatic vertebrate animal that is covered with scales, and equipped with two sets of paired fins and several unpaired fins.
# [fish] subClassOf [any aquatic vertebrate animal that is covered ...]

Construct {?sub rdfs:subClassOf ?super} {
  ?is a penn:BePresentTense .
  ?is nlp:superToken ?is_any_aquatic_.
  ?is_any_aquatic_ a olia:VerbPhrase .
  ?animal nlp:cop ?is .
}
Creating Knowledge out of Interlinked Data

Use Case - Machine Learning

Learned Concept

(Sentence and hasTaken some [ME and has lemma value "Essen"]

Acc.: 100%
This concept was created and saved by another user and saved as:
Sentences with Essen the city

Related concepts
hide

Concepts of other Users
hide
Sentences with Essen the city
load

Learning Input

start learning (needs 1 pos and 1 neg example)

Positive Samples

Da in den alten Ländern 1991 der Erdgasverbrauch um sieben Prozent auf die neue Rekordhöhe von 73,5
Millionen Tonnen Steinkohle einheiten wuchs, kündigten die Kassen in Essen noch lauter als im Jahr davor.

Die Idee stammt aus Dortmund und Essen. Auf ihrer dreitägigen Delegiertenversammlung in Essen...
Use Case - Machine Learning

Experiment 1
Passive with Participle and "werden"

- BASELINE
- ADAPT_X3
- REDUCE
- NO_LEMMA

Experiment 2
Passive with "zu" and Infinitive

- BASELINE
- ADD_10
- ADD_10_X3
- NO_LEMMA
Creating Knowledge out of Interlinked Data

Workplan

- EU Deliverable almost finished
- Integration of SnowballStemming and the Stanford Parser
- Next step: Integration of Knowledge Extraction tools (Zemanta, DBpedia Spotlight, Alchemy, OpenCalais, FOX)
- Web Service that read NIF and Output NIF
- Web Site: http://aksw.org/Projects/NIF
Summary

• NIF allows to represent NLP output using Knowledge Representation Formalisms (RDF/OWL)

• It is possible to mix it with other Knowledge (e.g. Wikipedia/DBpedia)

• Good foundation to optimize machine learning:
  • Choose the best algorithms
  • Choose the best data
Related Projects

- Wiktionary
- LLOD
- CKAN / Open Linguistics
### Creation of data sets: Wiktionary2RDF

<table>
<thead>
<tr>
<th>Language</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>English</strong></td>
<td><em>The free dictionary</em> 2,197,000+ entries</td>
</tr>
<tr>
<td><strong>Français</strong></td>
<td><em>Le dictionnaire libre</em> 1,915,000+ articles</td>
</tr>
<tr>
<td><strong>中文</strong></td>
<td>自由的多語言詞典 952,000+條詞條</td>
</tr>
<tr>
<td><strong>Malagasy</strong></td>
<td><em>Raki-bolana malalaka</em> 278,000+ teny</td>
</tr>
<tr>
<td><strong>Русский</strong></td>
<td>Свободный словарь 267,000+ статей</td>
</tr>
<tr>
<td><strong>Polski</strong></td>
<td><em>Wolny słownik</em> 201,000+ stron</td>
</tr>
<tr>
<td><strong>Lietuvių</strong></td>
<td><em>Laisvasis žodynas</em> 552,000+ straipsniai</td>
</tr>
<tr>
<td><strong>Türkçe</strong></td>
<td><em>Özgür sözlük</em> 270,000+ madde</td>
</tr>
<tr>
<td><strong>Tiếng Việt</strong></td>
<td>Từ điển mở 229,000+ mục từ</td>
</tr>
</tbody>
</table>

---

**Search & Languages**
- search • rechercher • 搜尋 • paieška • tadiavo • ara • поиск • tìm kiếm • szukaj • ശേഖരിക്കുന്ന • serchez • 찾기 • pesquisa • haku • keresés • ἀναζήτηση • volltext • sok • sök • ricerca • zoeken
Creation of data sets: Wiktionary2RDF

http://en.wiktionary.org/wiki/house

- Covers 170 languages
- Total of 10 million pages
- 900,000 users
- RDF Dump will increase number of editors
- Same properties as Wikipedia (stable identifiers)

- Hundreds of Wiktionary parsers (especially for English)
- Information is trapped in the Wiki
- Structure changes make software obsolete

Why try it again?

- DBpedia Extraction Framework is very mature (5 years, 15 developers)
- Configuration over Code, Templates will allow Wiktionarians to update Parsers
- Early contact with the community
Wiktionary, Wortschatz, OLiA can become the **Crystallization point** for a Linguistic Linked Data Web

Four major types:
- Lexical Semantic Resources
- Dictionaries
- Corpora
- Schemas/Ontologies
Open Licences – Focus of LOD2 and OKFN

CKAN is an open registry of data and content packages. Harnessing the CKAN software, this site makes it easy to find, share and reuse content and data, especially in ways that are machine automatable.

**Working Group on Open Data in Linguistics**
http://linguistics.okfn.org

- Founded on Nov 2010
- 40 Members
- Membership open, please join
- Over 100 data sets in CKAN
Thank you for your attention!